

**Appendix Contents**

<b>A Additional Related Works</b>	<b>14</b>
<b>B Additional Details of MULTIVERSE</b>	<b>14</b>
<b>C Further Analysis of MULTIVERSE</b>	<b>15</b>
C.1. Interaction Goal . . . . .	15
<b>D More examples on MULTIVERSE</b>	<b>17</b>
<b>E Prompt Templates used for MULTIVERSE</b>	<b>22</b>

## A. Additional Related Works

Recent advancements in LLMs have predominantly adopted simplistic yet highly effective architectures, notably through the model-stitching concept. Numerous prior studies have introduced various design modifications to bridge the performance gap with closed-source LLMs, such as GPT-4v [54], GPT-4o [18], Gemini-Pro [61], and Claude-3 family [1]. These efforts include focusing intently on high-resolution processing [40, 42, 58], implementing locality-enhanced projectors [7], and incorporating knowledge embeddings [29], layer traversal technique [28] and leveraging a diverse array of vision encoders [48, 63] have also been explored. Additionally, integrating external, task-specific computer vision modules [20, 27, 30, 31] and incorporating different modalities — including video and audio [13, 38, 65, 69] — have expanded the models’ capabilities. Moreover, enabling the handling of interleaved input formats [39, 70] has further broadened the versatility of these models.

## B. Additional Details of MULTIVERSE

Datasets	# Images	Public Links	Licenses
MMDU [47]	375	<a href="https://huggingface.co/datasets/laolao77/MMDU">https://huggingface.co/datasets/laolao77/MMDU</a>	CC-BY-NC-4.0
MegaBench [8]	12091	<a href="https://huggingface.co/datasets/TIGER-Lab/MEGA-Bench">https://huggingface.co/datasets/TIGER-Lab/MEGA-Bench</a>	Apache License 2.0
GroundUI-1K [79]	1000	<a href="https://huggingface.co/datasets/agent-studio/GroundUI-1K">https://huggingface.co/datasets/agent-studio/GroundUI-1K</a>	MIT
MMMU [73]	12141	<a href="https://huggingface.co/datasets/MMMU/MMMU">https://huggingface.co/datasets/MMMU/MMMU</a>	Apache License 2.0
MMMU-Pro [74]	2015	<a href="https://huggingface.co/datasets/MMMU/MMMU_Pro">https://huggingface.co/datasets/MMMU/MMMU_Pro</a>	Apache License 2.0
NaturalBench [37]	3800	<a href="https://huggingface.co/datasets/BaiqiL/NaturalBench">https://huggingface.co/datasets/BaiqiL/NaturalBench</a>	Apache License 2.0
VisIT-Bench [6]	574	<a href="https://huggingface.co/datasets/mlfoundations/VisIT-Bench">https://huggingface.co/datasets/mlfoundations/VisIT-Bench</a>	CC-BY-4.0
MathVista [50]	6141	<a href="https://huggingface.co/datasets/AI4Math/MathVista">https://huggingface.co/datasets/AI4Math/MathVista</a>	CC-BY-SA-4.0
MM-Vet [71]	218	<a href="https://huggingface.co/datasets/whyu/mm-vet">https://huggingface.co/datasets/whyu/mm-vet</a>	CC-BY-NC-4.0
MM-Vet v2 [72]	704	<a href="https://huggingface.co/datasets/whyu/mm-vet-v2">https://huggingface.co/datasets/whyu/mm-vet-v2</a>	CC-BY-NC-4.0
CharXiv [67]	2323	<a href="https://huggingface.co/datasets/princeton-nlp/CharXiv">https://huggingface.co/datasets/princeton-nlp/CharXiv</a>	CC-BY-SA-4.0
MMBench (Eng) [46]	6718	<a href="https://huggingface.co/datasets/lmms-lab/MMBench_EN">https://huggingface.co/datasets/lmms-lab/MMBench_EN</a>	Apache License 2.0
<b>All Images</b>	<b>48100</b>	-	-

Table 4. Detailed information on the 16 source datasets collected for MULTIVERSE.

In Step 1, we collect 55,500 images from 16 existing evaluation benchmarks. The detailed statistics, including the number of images, public links, and license information, are provided in Table 4.

**License.** All evaluation benchmarks permit non-commercial use, allowing us to preprocess and redistribute the data. However, certain datasets, such as MathVista and CharXiv, are licensed under CC-BY-SA-4.0. This requires us to comply with the same licensing terms when publicly deploying MULTIVERSE.

## C. Further Analysis of MULTIVERSE

### C.1. Interaction Goal

**Verification (87)** : Data Validation, Data Consistency, Statistical Significance, Financial Verification, Graph Intersection Points, Cost Accounting Verification, Strategy Validation, Equation and Calculation Verification, Answer Verification, Image Details Verification, Term Definition Matching, Mathematical Theorem and Calculation Verification, Graph Interpretations, Verification of Calculations and Materials, Anatomical Details Verification, Molecular Structure and Properties, Design Verification, Mechanical Dynamics Calculations, Image Verification, Graph Interpretation, Sequence Evaluation, Dimension Verification, Financial Viability, Biopsy Image Analysis, Geometric Problem Verification, Feedback Control System Design, Data Interpretation Verification, Data Accuracy Verification, Equation Verification, Data Interpretation, Structural Calculations, Feedback, Musical Score Interpretation, Location and Meaning, Financial Analysis, Code Verification, Mathematical Properties, Solution Verification, Data Correlation, Graph Theory, Model Interpretation Verification, Interpretation, Alignment with Theoretical Models, Interpretation Verification, Geometric Properties, Molecular Structure Verification, Optimization, Geometry, Compatibility Verification, Geometric Verification, Data Accuracy, Audio Issue Resolution, Verifying Distribution Shape, Logic Verification, Calculation Verification, Data Analysis, Theoretical Implications, Identification, Music Theory, Theoretical Consistency, Molecular Structure, Mathematical Method, Structural Analysis Calculations, Safety and Performance Evaluation, Circuit Calculation Verification, Circuit Verification, Function Identification, Formula Verification, Algorithm Verification, Molecular Structure Confirmation, Feedback Collection, Data Verification, Financial Projections, Confirmation, Enhancement, Mathematical Verification, Feedback Assessment, Parameter Suitability, Diagnosis, Solution Checking, Signal Analysis Verification, Cost Verification, Interpretation Checking, Design Feedback, Thermodynamic Properties, Financial Statement Compliance, Accuracy Verification

**Analysis (123)** : Problem Solving, Chess Endgame Strategy, Weather Patterns, Model Convergence in Option Pricing, Mechanical Stress Distribution, Stress Analysis, Art Analysis, Pattern Recognition, Data Comparison, Trend Analysis, Pattern Detection, Impact of Sparsity on Data Recovery, Disease Outbreak Factors, Risk-Return Trade-offs, Schedule Optimization, Connectivity Analysis, Graph Data Interpretation, Statistical Methods Comparison, Travel Data Analysis, Impact Assessment, Technique Evaluation, Performance Metrics, Performance Comparison, Classification, Music Analysis, Health Disparities, Investment Strategy Analysis, Cost and Time Implication Analysis, Cost Optimization, Model Interpretation, Economic Analysis, Chess Endgame Analysis, Feedback Evaluation, Data Interpretation, Probability Distribution Models, Performance Interpretation, Performance Analysis, Bias Detection, Design Elements Assessment, Financial Analysis, Game Evaluation, Data Pattern Interpretation, Architectural Analysis, Economic Impact, Model Performance Evaluation, Geometric Pattern Analysis, Mechanical Forces, Diagnosing Plant Issues, Strategic Planning, Artistic Elements and Symbolism in Manga, Design Evaluation, Material Evaluation, Price Comparison, Trade-offs Assessment, Measurement Suitability, Causation Analysis, Interpretation, Geometric Relationships Analysis, Traffic and Urban Design, Investment Options, System Behavior Analysis, Behavioral Analysis and Health Assessment, Medical Condition Assessment, Project Management Insights, Symbolism Identification, Equity Changes, Force Distribution Study, Property Interpretation, Artistic Analysis, Strategy Evaluation, Risk-Return Trade-off, Ecosystem Health Assessment, Demographic Analysis, Financial Health Analysis, Data Analysis, Decision-making Analysis, Identification, Correlation Analysis, Risk and Return, Musical Analysis, Suitability Assessment, Design Elements, Trend Identification, Curve Identification, Strategic Analysis, Optimization Analysis, Statistical Analysis, Memory Management, Market Research, Function Identification, Spectral Analysis, Electricity Usage Analysis, Comparative Analysis, Performance and Robustness Evaluation, Enhancing Engagement Strategies, Summarization, Nuclear Reaction Analysis, Comparative Evaluation, Data Trends, Chess Strategy, Quantum Mechanics, Function Behavior Analysis, Financial Market Analysis, Cause Identification, System Stability, Research Analysis, Diagnosis, Empirical vs Theoretical Comparison, Investment Strategy Performance, Damage Assessment, Sales Trend Analysis, Medical Analysis, Infection Spread Dynamics, Interpretation of Experimental Data, Pattern Analysis, Sustainability, Visual Techniques, Comparison, Performance Metrics Interpretation, Data Visualization, Feedback on Composition, Correlation Study, Stock Market Trends

**Exploration (30)** : Instructional Strategies, Idea Generation, Cultural and Historical Exploration, Geometric Properties, Design Elements, Information Gathering, Cultural and Historical Insights, Music Analysis, Ingredient Identification, Music Discovery, Weather and Photography, Geometric Structure Identification, Applications of Geometric Art in Architecture, Design Improvements, Algorithm Development, Species Identification, Model Recommendations, Design and Aesthetics, Comparison and Analysis of Options, Puzzle Solving Strategies, Interpretation of Visual Elements, Feature Discovery, Species

Recognition, Assessment of Urban Environmental Features, Design Variations, Art and Design Inspiration, Chemical Compound Identification, Information Discovery, Creative Applications, Coordinate Conversion

**Optimization (48)** : Process Improvement, Readability Enhancement, Algorithm Efficiency, Strategic Improvement, Project Management, Scheduling, Material Selection, Code Quality, Design Efficiency, Pathfinding, Translation Model Performance Enhancement, Improving Visual Appearance, Gameplay Strategy, Code Efficiency, Neural Network Configuration, Team Building, Cost Optimization, Image Enhancement, Inventory Management, Visualization Clarity, Decision Making, Route Efficiency, Rendering Techniques Improvement, Strategy Development, Code Optimization, Process Optimization, Eco-friendliness, Investment Portfolio, Clarity Enhancement, Pricing Strategy Optimization, Integration of auditory and visual elements, Schedule Optimization, Resource Allocation, Enhancing Real Estate Listings, Visual Design, System Efficiency Enhancement, Strategic Planning, Code Efficiency and Readability, Price Comparison, Resource Management, Workflow Efficiency, Organization, Improve Data Visualization, Script Reliability, Data Visualization, Robot Movement Strategy, Efficiency Enhancement, Vibration Isolation Methods

**Calculation (32)** : Financial Calculation, Problem Solving, Arithmetic Problem, Project Management, Geometry Problem Solving, Geometric Calculation, Investment Evaluation, Geometry, Financial Computation, Structural Analysis, Force Calculation, Finance, Cost Analysis, Decision Making, Investment Returns, Weighted Average Calculation, Pressure Difference Calculation, Geometry Calculation, Financial Analysis, Investment Performance Analysis, Equilibrium Calculation, Trajectory Calculation, Load Analysis, Financial Modeling, Statistical Metric Calculation, Solving Equations, Interpolation, Structural Engineering, Cost Estimation, Stress Intensity Factors, Trigonometry, Geometrical Calculation

**Understanding (86)** : Concept Clarification, Mechanics, Comprehension of Data Visualizations, Technical Concepts, Enrollment Disparities, Chemical Properties and Applications, Scientific Concepts, Model Performance, Implications of OOD detection in medical predictions, System Settings, Electrical Circuit Analysis, AI Capabilities, Mathematical Relationships, Material Impact, Comprehension of Concepts, Comprehension of Biological Systems, Comprehension of Graph Data, Thermodynamics Concepts, Solution Process, Educational Trends, Learning how to create geometric patterns, Conceptual Differences, Graph Interpretation, Deciphering, Structural Mechanics, Effects of Data Augmentation, Code Logic and Concepts, Economic Concepts, Behavior and Applications, Function Behavior, Seismic Data Interpretation, Simplification, Advanced Mathematics Problems, Understanding Electrostatic Potential, Data Interpretation, Performance Analysis, System Dynamics, Algorithm Logic Understanding, Comprehension of Construction Techniques, Ecological Roles, Explaining Properties and Applications, Comprehension of Theoretical Concepts, Graph Theory, Structural Engineering, Application of Frameworks, Automotive Dynamics, Historical Context, Geometric Properties, Historical and Cultural Analysis, Concept Explanation, Function and Importance of DVD Layers, Research Methodologies, Functionality Explanation, Plant Growth and Reproduction, Technology Applications, Symbolism Analysis, Sales Insights, Understanding Concurrency Issues, Music Theory Comprehension, Clarification, Historical and Architectural Significance, Geometric Concepts, Biology Concepts, Music Theory, Symbolism in Art, Application of Technology, Color Interaction and Emotional Impact, Circuit Configuration and Components, Cultural Significance, Conceptual Clarification, Conceptual Comprehension, Comparative Analysis, Comprehension of Map Projections, Comprehension of Tree Traversals, Code Structure and Functionality, Chemical Reactions, Behavior Analysis, Reinforcement Learning Policies, Conceptual Understanding, Terminology, Relationship Analysis, Clarification of Processes, Leadership Pyramid Levels, Algorithmic Optimization, Color Theory, Comprehension of Dynamics

**Research (72)** : Data Retrieval, Historical and Route Information Retrieval, Teaching Techniques, Design Insights Gathering, Recipe Acquisition, Historical Information Gathering, Gathering Resources, Information Gathering, Product Identification and Supplier Sourcing, Historical and Architectural Research, Hotel Exploration, Historical Significance, Gathering Information, Historical and Architectural Information, Educational Resources, Economic Trends Analysis, Art Analysis, Historical and Technical Information, Cultural Studies, Information Retrieval, Biological Research, Trend Analysis, Data Interpretation, Species Identification, Art History, Architectural Patterns, Travel Logistics, Cultural and Historical Analysis, Trends Analysis, Historical Insights, Product Information Gathering, Safety Regulations and Design Practices, Ecological Study, Literature Review, Sports Analysis, Identification and Knowledge Acquisition, Art, Historical and Mythological Context, Ecological Information Gathering, Cultural and Historical Research, Film Analysis, Historical and Architectural Details, Gathering Expert Tips, Product Features and Benefits, Identification, Historical Research, Art Style Similarity, Symbiosis in Marine Biology, Design Origins, Historical Analysis, Horticulture, Mentorship and Collaboration Opportunities,



Historical Events, Sustainable Design Insights, Historical and Architectural Analysis, Disease Identification and Management, Geological Study, Art Techniques, Historical Economic Impact, Comparison, Art and Artist Analysis, Art Historical Contextualization, Geology, Agricultural Knowledge Gathering, Historical Context, Historical Design Analysis, Cultural and Historical Facts Gathering, Ecology, Historical and Cultural Significance, Collaboration, Music History, Architectural Insights

**Creation (43)** : Explain, Educational Activity Design, Interior Design, Creative Cooking, Design Brainstorming, Design Enhancement, Lyric Writing, Chart Creation, Infographic Design, Lesson Plan Design, Word Association, Design Inspiration, Content Generation, Activity Design, Schedule Design, Design Integration, Developing Photography Exhibit Concepts, Educational Content Development, Interactive Educational Activities Design, Guide Development, Design Generation, Question Development, Content Creation, Educational Material Development, Exercise Development, Data Compilation, Digital Transcription, Discussion Generation, Humor Enhancement, Culinary Innovation, Culinary Creation, Material Enhancement, Recipe Development, Creative Ideation, Cocktail Ingredient Arrangement, Data Visualization, Demonstration Development, Game Development, Data Visualization Enhancement, Engagement, Culinary Innovating, Visualization, Artistic Inspiration and Design

#### **D. More examples on MULTIVERSE**

We present more examples on MULTIVERSE in [Figure 10](#), [Figure 11](#), [Figure 12](#), and [Figure 13](#).

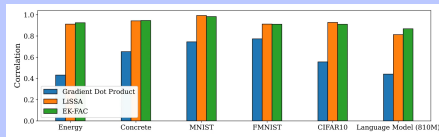


Figure 6: Performance comparison of the gradient dot product, LiSSA, and EK-FAC influence estimation methods as measured by Pearson correlation with the PBRF. The correlations were averaged over 10 measurements, and 500 training data points were used to measure the correlation. EK-FAC outperforms the gradient dot product and achieves performance comparable to LiSSA across all tasks.

I've been analyzing these influence estimation methods. Can you explain how they compare overall?

The image shows a comparative analysis of three influence estimation methods: Gradient Dot Product, LiSSA, and EK-FAC, evaluated using Pearson correlation with a specific benchmark. Across various tasks such as Energy, Concrete, MNIST, FMNIST, CIFAR10, and a large language model, EK-FAC consistently outperforms the other methods. It not only achieves higher correlation values but also exhibits robust performance across diverse datasets. The Gradient Dot Product, while useful, lags behind both LiSSA and EK-FAC, especially in large-scale tasks indicated by the Language Model dataset, where its correlation is notably lower.

Interesting. Can we dive deeper into why EK-FAC shows superior performance? Is it related to its computational efficiency or data handling?

EK-FAC's superior performance is likely attributable to both its computational efficiency and its innovative data handling capabilities. Firstly, EK-FAC uses efficient Kronecker-factored approximations, which allow for faster updates and scalability, vital in handling large datasets or models such as the ones indicated in the Language Model results.

Conceptually, EK-FAC approximates the Fisher information matrix more effectively, leading to more accurate influence estimates. This precision is critical in understanding how different data points affect model predictions across diverse datasets. Additionally, when it comes to computational efficiency, EK-FAC reduces the complexity associated with matrix operations, making it a practical choice for both small- and large-scale applications, as demonstrated in the comparative results.

Could you provide potential challenges or limitations associated with EK-FAC that I should consider while preparing for the conference?

Certainly. While EK-FAC shows promising results, there are several considerations and potential challenges. First, the implementation complexity is higher compared to simpler methods like the Gradient Dot Product. This added complexity requires careful tuning and understanding of the approximations used.

Additionally, while EK-FAC is computationally efficient for large models, the initial setup and hyperparameter tuning could be more time-consuming, potentially impacting performance if not done correctly. There might also be limitations related to the approximation's fidelity in extreme cases or specific data distributions not well-represented in the showcased tasks.

Furthermore, since EK-FAC is relatively advanced, it might present integration challenges within existing systems or require significant adaptation to align with varied frameworks, something important to highlight during your presentation.

Given the diverse datasets, how might these methods perform under adversarial conditions or data perturbations?

Under adversarial conditions or data perturbations, the performance of these influence estimation methods can significantly vary. EK-FAC's robust approach might offer an advantage given its sophisticated handling of data dependencies and inner workings of model structures; however, even it is not immune to well-crafted adversarial attacks.

LiSSA can also offer resilience due to its iterative nature, which might help in recalibrating model understanding when encountering anomalies. Gradient Dot Product, while more straightforward, might be more susceptible to adversarial noise due to its reliance on basic gradient information without deeper model insight.

Enhancing their performance under such conditions often involves integrating additional defense mechanisms, like adversarial training or noise-resistant layers, especially in contexts where model security is a priority. These are critical elements to discuss to address potential vulnerabilities and bolster robustness in your presentation.

Figure 10. An example of MULTIVERSE.



I came across this iconic film listing and started revisiting the differences between the book and the movie. Can you share any intriguing production trivia related to it?

Certainly! "Harry Potter and the Chamber of Secrets" faced several exciting production challenges and moments. One notable piece of trivia is the filming of the scene in the Chamber itself, which was constructed with intricate sets to represent the ominous atmosphere of Salazar Slytherin's lair. The filmmakers used over 70 live snakes to enhance the realism of the serpentine environment.

Another interesting tidbit involves Dobby the house-elf, who was entirely a CGI creation. To aid actors during scenes with Dobby, a tennis ball on a stick or a puppet was used for eyelines. This led to spontaneous and memorable improvisations, especially for reactions and interactions.

Additionally, the Whomping Willow's effects team made a model tree that could be manipulated hydraulically, allowing for believable and dynamic shots of it thrashing around and causing mayhem, especially during the infamous flying car scene.

Fascinating! How about some significant changes or omissions in the storyline compared to the book? What should I note for my blog?

There are several key differences between the book and the movie adaptation. One prominent change is the absence of the Deathday Party. In the book, Nearly Headless Nick invites Harry, Ron, and Hermione to this ghostly gathering, offering a deeper glimpse into the ghostly community at Hogwarts. This scene was omitted from the film for pacing reasons.

Additionally, the book gives more extensive insight into character backstories, such as the history of Hagrid's time at Hogwarts and his expulsion, which is covered only briefly in the movie.

The dialogue with Tom Riddle during the climax in the Chamber is also more expanded in the book, providing greater detail into his past and relationship with Harry. This intricacy adds layers to the characterization of Voldemort and his connection to Harry, setting a richer backdrop for future conflicts.

Now, looking deeper, can you help clarify how thematic elements were altered or emphasized between the two mediums?

Absolutely! One thematic element that holds prominence is the concept of identity and heritage. The book delves deeply into the theme of identity through the lens of blood purity, a frequent topic in the wizarding world, with particular emphasis on the derogation of 'Mudbloods.' This theme is encapsulated in personal struggles, especially for Hermione. While the movie touches on these themes, the book offers a more profound critique of prejudice and the moral implications surrounding it.

Moreover, the book explicitly explores the fear and curiosity surrounding the mysterious Tom Riddle's diary, emphasizing Harry's internal conflict and questions about his own identity and abilities. This psychological tension is slightly streamlined in the film to fit the narrative structure and keep the pacing brisk, focusing more on the thrilling aspects of the storyline.

The film does well to evoke the sense of adventure and suspense visually, through dazzling special effects that bring creatures like Aragog and the Basilisk to life, yet it condenses some of the introspective narrative elements that explore character depth and moral dilemmas.

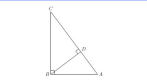
Lastly, could you draft a concise comparison paragraph that summarizes these insights for my blog? Something engaging yet informative.

Certainly! The adaptation of "Harry Potter and the Chamber of Secrets" from book to screen is a mesmerizing yet nuanced transition. The film accentuates its visual grandeur and suspense-driven narrative, manifested through incredible set designs and groundbreaking CGI techniques, notably in scenes involving Fawkes the phoenix and Dobby the house-elf. Conversely, the novel more thoroughly examines themes of identity and heritage, delving into the contentious dialogues on blood purity and offering a poignant exploration of character backstories, such as Hagrid's troubled history. This creates a layered narrative that challenges readers to reflect on broader societal issues, showcasing the complexity of J.K. Rowling's magical world—elements that are subtly abridged in the cinematic version to maintain narrative fluidity. Observing these variations offers a unique lens for understanding how each medium brings its strengths to this enchanting tale.

Figure 11. An example of MULTIVERSE.



Figure 12. An example of MULTIVERSE.



6. In the figure above,  $BD = 6$  and  $BC = 8$ . What is the length of segment  $BD$ ?

(A) 3  
 (B)  $\frac{24}{5}$   
 (C) 4  
 (D)  $\frac{12}{5}$   
 (E) 8

I've been trying to solve this problem, and I think I got it, but I'm not completely sure. Could you help me verify my solution for finding the length of one of the segments here?

Certainly! Given that you are trying to find the length of segment  $BD$  in this right triangle with provided lengths  $AB = 6$  and  $BC = 8$ , we'll use the properties of right triangles. Specifically, we can apply the Pythagorean theorem to the right triangle  $ABC$ .

First, let's calculate  $AC$  using the Pythagorean theorem:

$$AC^2 = AB^2 + BC^2 = 6^2 + 8^2 = 36 + 64 = 100$$

Thus,  $AC = \sqrt{100} = 10$ .

Now, we need to find the length of  $BD$ . Since  $BD$  is an altitude to the hypotenuse  $AC$  in triangle  $ABC$ , we can use the relation  $BD^2 = AB \times BC \div AC$  considering that it's a property of similar triangles formed here in a right triangle:

$$BD = \frac{AB \times BC}{AC} = \frac{6 \times 8}{10} = 4.8$$

However, you should check if this value corresponds to one of the multiple-choice options. The closest fraction that represents this value is  $\frac{24}{5}$ .

That makes sense now. What if the triangle had a different orientation or the line was in a different position? Would the approach change?

That's a great question! If the triangle's sides are altered or rotated, the foundational geometry principles would still apply, but the specific triangles used for analysis might differ. In the case of a right triangle, where an altitude is drawn to the hypotenuse, the relationships are consistent irrespective of orientation.

If  $BD$  is an altitude to the hypotenuse, the formula  $BD^2 = AB \times BC \div AC$  remains valid. This stems from the property of the similarity of triangles: each smaller triangle formed by the altitude is similar to the original triangle, preserving proportionate relationships even with a varied orientation.

For different geometries or altitudes not drawn to the hypotenuse, we'd need alternative strategies based on triangle properties applicable in each scenario, such as the Law of Sines or Cosines in non-right triangles.

Thanks! What if there were additional constructs or angles given in this problem? How would I tackle that information?

Additional constructs, such as medians, bisectors, or angle measures, enrich the problem with more geometric nuances, bringing a broader range of theorems and properties into play.

For instance, if angles were given or required, angle bisector theorems or trigonometric relationships might be necessary. Medians involve centroids, where you consider dividing medians into segments in a 2:1 ratio, and bisectors allow application of the Angle Bisector Theorem to find proportional lengths.

If circles (e.g., incircles, circumcircles) are involved, properties related to tangents or inscribed angles would also come into play. It's essential to identify which geometric principles relate directly to the extra elements and systematically explore each construct's mathematical implications within the triangle's framework.

That's a comprehensive overview! Lastly, could you explain how to use coordinate geometry to verify these calculations if the triangle's vertices were given on a plane?

Using coordinate geometry is a powerful method to authenticate solutions with precision by leveraging geometric coordinates.

- Positioning Points:** Assume  $B = (0, 0)$ ,  $A = (6, 0)$ , and let  $C$  be the point to satisfy the triangle's conditions. Given  $BC = 8$ , the point is along the  $y$ -axis, i.e.,  $C = (0, 8)$ .
- Verifying Altitude:** With these coordinates, the line  $BD$  as an altitude implies it is perpendicular to  $AC$ . The coordinates for  $D$  can be obtained by solving equations based on the slopes being negative reciprocals for perpendicular lines, or use the distance properties directly.
- Calculating Lengths:** Apply the distance formula:
 
$$BD = \frac{|C_x + D_y + E|}{\sqrt{A^2 + B^2}}$$
 Substituting values:
 
$$BD = \frac{|0 \cdot 6 + 1 \cdot 0 + (-8)|}{\sqrt{1^2 + (-6/8)^2}} \approx \frac{8}{\sqrt{1.5625}} = \frac{8}{1.25} = 4.8$$

These calculations confirm previous results and illustrate the versatility of verifying solutions through multiple geometric lenses.

Figure 13. An example of MULTIVERSE.

## E. Prompt Templates used for MULTIVERSE

### Prompt Template for Scoring Image Quality

You are given an image. Your task is to evaluate the image quality, which is described as follows:

[Image Quality Evaluation]

Rate the given image on a scale from 1 to 5 based on its quality, considering image clarity, resolution, and likelihood of occurrence in real-world conversations:

- Score 1 (Very Low Quality): The image lacks clarity, and objects or content appear blurry with minimal detail, making it unlikely to appear in real conversations. Additionally, this score applies if the image includes a specific brand logo or icon or if the image itself has been rotated.
- Score 2 (Low Quality): The image is not sharp, with only some objects or elements visible, and is rarely used in real conversations.
- Score 3 (Moderate Quality): The image is moderately clear in parts, with sufficient visibility of objects or content to be recognized in specific situations or by certain individuals in conversation.
- Score 4 (High Quality): The image is clear, with most objects or elements recognizable, and is likely to appear in real conversations.
- Score 5 (Very High Quality): The image is extremely clear, with even small objects or details highly visible, making it highly relevant for frequent use in real conversations.

Please generate your answer by strictly following the guidelines below:

[Guidelines]

- The answer should be formatted as a Python dictionary containing the following key: "image\_quality\_score".
- The "image\_quality\_score" should indicate the quality of the given image, range from 1 to 5.

[Answer]

### Prompt Template for Image Category

You are given an image. Your task is to assign a subject/category to the image, which is described as follows:

[Image Category Classification]

Classify the subject/category of the given image based on its content. Choose from the following categories, along with the corresponding descriptions:

- Vehicles and Transportation: Includes all types of transportation, such as cars, bikes, trains, planes, and ships.
- Food and Cuisine: Covers cuisine styles, dishes, ingredients, and cooking activities.
- People and Lifestyle: Encompasses everyday life, cultural lifestyles, and social activities involving people.
- Sports and Recreation: Includes all sports, games, and leisure activities.
- Animals and Wildlife: Covers both domestic animals and wildlife, with subcategories like birds, sea creatures, etc.
- Objects, Clothing, and Accessories: Encompasses fashion, personal items, and any identifiable object.
- Brands and Products: Relates to consumer goods, brand logos, and popular products.
- Architecture and Landmarks: Includes buildings, bridges, monuments, and notable geographic features.
- Tradition and History: Covers historical sites, artifacts, and traditional practices.
- Fine Art and Illustrations: Includes classical paintings, digital art, and creative illustrations.
- Celebrities and Public Figures: Specific to well-known individuals in popular culture or politics.
- Science and Technology: Includes scientific equipment, laboratories, and research environments.
- Chemistry and Lab Equipment: Covers chemicals, laboratory settings, and molecular structures.
- Mathematics and Symbols: Encompasses mathematical diagrams, equations, and geometric shapes.
- Nature and Landscapes: Includes scenery, forests, mountains, rivers, and natural phenomena.

- Healthcare and Medicine: Covers medical devices, doctors, hospitals, and health-related content.
- Programming and Coding: Depicts code, software interfaces, and development environments.
- Web Design and User Interfaces: Encompasses UI/UX layouts, website mockups, and interface elements.
- Mobile and Smart Devices: Includes smartphones, tablets, and other handheld devices.
- Weather and Climate: Covers weather events, climate patterns, and related data visualization.
- Festivals and Events: Encompasses celebrations, public gatherings, and cultural events.
- Education and Learning: Covers classrooms, books, e-learning visuals, and study materials.
- Entertainment and Media: Includes visuals from movies, shows, music, and gaming.
- Interior Design and Furniture: Relates to home decor, furniture styles, and interior layouts.
- Natural Disasters and Environmental Hazards: Includes earthquakes, floods, pollution, and environmental issues.
- Abstract and Conceptual Art: Covers non-representational or symbolic images.
- Business and Finance: Depicts financial graphs, business meetings, currency, and economy-related visuals.
- Law and Government: Includes courts, governmental symbols, legal documentation, and policy.
- Diagrams and Schematics: Covers flowcharts, technical diagrams, mind maps, and other schematic representations.
- Charts and Graphs: Encompasses data visualizations like bar charts, pie charts, line graphs, and other statistical representations.
- Documents and Paperwork: Covers text-heavy documents, official forms, written notes, and reports.
- Receipts and Invoices: Includes financial documents, bills, purchase receipts, and transaction records.
- Others

Please generate your answer by strictly following the guidelines below:

[Guidelines]

- The answer should be formatted as a Python dictionary containing the following key: “image\_category”, and “image\_sub\_category”.
- The “image\_category” should contain the most appropriate classified category of the given image.
- The “image\_sub\_category” should specify a more detailed sub-category within the selected “image\_category” to provide a finer level of classification.
- If you select “Others”, then please generate a new image category.

[Answer]

### Prompt Template for Personal Background Generation

You will be provided with an image.

Your task is to:

#### 1. Fictional Character Creation

- You should create a fictional character who plausibly engages in a conversational interface with an AI assistant by providing a given image.
- The fictional character should initiate a realistic and contextually appropriate conversation based on the image.
- The fictional character should be represented in 2-3 sentences (NOT a structured format), covering aspects like name, age, gender, occupation, preference, hobby, like/dislike, or background knowledge (e.g., basic, intermediate, expert).
- If there is a real human in the given image, please do NOT generate the fictional character corresponding to the real human in the given image. You MUST generate a new fictional character who will share the given image with the AI assistant in the conversational interface.

#### 2. Scenario Context Creation

- You should create a plausible and detailed scenario in which the fictional character interacts with the AI assistant about the given image through a conversational interface.
- The scenario context should realistically unfold in a multi-turn conversation between the fictional character and the AI assistant.
- The scenario should be practical, realistic, and engaging, reflecting real-world situations.

### 3. Goal Creation

- You should create a specific and realistic goal that the fictional character has when engaging in a conversation with the AI assistant through a conversational interface by providing a given image, in alignment with the generated scenario context.
- The goal should be relevant to the scenario context and represented as a concise phrase. For example, the assistant might search for information, retrieve data, solve a math or coding problem, etc.

Please brainstorm the most appropriate, realistic, and highly plausible fictional character, scenario context, and goal that could naturally occur in a real-world conversation based on the given image by strictly following the JSON format below:

- The answer should be formatted as a JSON-formatted Python dictionary containing the following keys: “Character Description”, “Scenario”, and “Goal”.

Answer:

## Prompt Template for Multi-Turn Conversation Generation

You will be provided with an image.

Your task is to brainstorm a creative, realistic and practical multi-turn conversation between a fictional character (denoted as [Fictional Character]) and the AI assistant, based on a given scenario context (denoted as [Scenario Context]) and the fictional character’s goal (denoted as [Goal]). In the conversation, the fictional character interacts with the AI assistant through a conversational interface by providing an image.

The fictional character, scenario context, and goal are presented as follows:

[Fictional Character]  
{character}

[Scenario Context]  
{scenario}

[Goal]  
{goal}

Please brainstorm a practical and realistic multi-turn conversation based on the given fictional character, scenario context, goal, and the provided image by strictly following the guidelines below:

### [Guidelines]

- The generated multi-turn conversation should be presented in plain text, where the fictional character’s utterances MUST begin with “USER:”, and the AI assistant’s utterances MUST begin with “ASSISTANT:”.
- As the conversation progresses, the fictional character should ask progressively more challenging, creative, complex requests, such as follow-up questions, knowledge acquisition, refinement, style/content rephrasing, advanced reasoning expansion, etc.
- The fictional character’s utterances should be creative, reflect diverse linguistic styles, and include questions that



require correct answers to difficult problems.

- The AI assistant's utterances **MUST** be very long detailed, specific, expert, helpful, and informative in addressing the fictional character's requests.
- The AI assistant's utterances should demonstrate advanced cognitive reasoning and do **NOT** include the fictional character's name.
- If the AI assistant's utterances involve code, mathematical equations, charts, tables, graphs, scientific concepts, API functions, planning, or any higher-order knowledge, they **MUST** be highly informative, expert, specific, detailed, and factual. They should also include clear and comprehensive explanations alongside the technical details (e.g., code, mathematical theorems, API function calls, etc.).
- Each turn consists of two utterances: one from the fictional character and one from the AI assistant.
- The conversation should not end with a closing remark like "See you next time" or similar.
- The conversation must consist of four turns (eight utterances), with each turn involving one utterance from the fictional character and one from the AI assistant.

Note:

- **NEVER** generate utterances from characters in cases where the AI assistant can resolve them precisely without viewing the given image.
- **ALWAYS** refer to the given image using pronouns (e.g., "it," "them") in any fictional character's utterances.
- **NEVER** explicitly mention the key entity or information depicted in the image in any fictional character's utterances.

[Generated Multi-Turn Conversation]

### Prompt Template for Checklist Generation

You will be provided with an image and the user's utterances from a multi-turn conversation. The conversation takes place in a conversational interface where the user and the AI assistant interact while referring to the given image.

Your task is to create an instance-specific evaluation checklist that will be used to assess the quality of the AI assistant's response to the user's query within the multi-turn conversation. In other words, you need to create the evaluation criteria (in a question format) that the AI assistant's response must meet to be considered the optimal response for the given user query, based on the given image.

The checklist should include multiple questions, each satisfying the following conditions:

- Each question must be answerable with "Yes" or "No."
- Each question must be relevant to one specific evaluation criteria from the provided criteria collection.
- Each question should minimize subjectivity in the rater's judgment.
- Questions should be formulated so that a "Yes" answer is a positive evaluation.
- Each question must evaluate aspects related to the conditions that need to be met—by referring to the given image and each turn of the user's query—to produce the optimal response, and these conditions must be strictly satisfied.

#### ### Evaluation Criteria Collection

Each evaluation criteria represents a main capability to evaluate (alongside the definition):

- **Engagingness:** Measures the model's ability to sustain an engaging and interactive conversation by assessing flow, immersion, interactivity, and emotional connection with the user.
- **Tone & Style Appropriateness:** Evaluates whether the response maintains an appropriate, positive, polite, and respectful tone.
- **Contextual Understanding:** Assesses how well the model understands the previous conversational context, including anaphora resolution and maintaining consistency across multiple turns.
- **Memory:** Evaluates whether the model accurately remembers and incorporates earlier dialogue details while tracking and retaining long-term contextual dependencies.
- **Proactiveness:** Determines whether the model proactively identifies and fulfills user needs by making helpful

suggestions and guiding the conversation.

- Clarity & Logical Structure: Evaluates the clarity, coherence, and logical organization of responses, ensuring they are easy to understand.
- Coherence: Ensures responses maintain consistency across turns and that dialogue progresses naturally without contradictions.
- Knowledge Understanding: Determines whether the model demonstrates expertise and incorporates relevant domain knowledge with depth and accuracy.
- Factual Correctness: Ensures responses contain accurate, verifiable facts and do not generate misinformation.
- Specificity & Informativeness: Evaluates how specific and detailed the response is, measuring the amount of useful, non-generic information provided.
- Cognitive Reasoning: Assesses the model's ability to reason logically and make commonsense inferences, including problem-solving and inference-making.
- Creativity: Evaluates the model's ability to generate creative and original content in storytelling, idea generation, coding, and visual responses.
- Problem-Solving Capability: Measures how effectively the model breaks down and addresses complex user queries, including step-by-step explanations.
- Helpfulness: Determines whether the response directly addresses user needs and provides actionable suggestions.
- Instruction Following: Assesses the model's ability to interpret and adhere to explicit user instructions without deviation.
- Harmlessness & Ethical Awareness: Ensures responses avoid harmful, offensive, or biased content while adhering to ethical considerations.
- Perceptual Understanding: Evaluates how well the model interprets the given image and integrates it into coherent responses.
- Fluency & Grammatical Accuracy: Assesses the linguistic correctness of responses, including grammar, spelling, sentence completeness, and verb tense consistency.
- Adaptability: Measures how well the model adapts responses based on user background, preferences, and conversational history.
- Multimodal Consistency: Ensures the model maintains semantic alignment between different modalities (e.g., text and images) and that textual responses accurately reflect visual inputs.

### ### Multi-Turn Conversation

Below is the multi-turn conversation (the AI assistant's responses are regarded as the optimal answers to the user queries). The conversation always starts at turn number 1.

{conversation}

—

Please brainstorm comprehensive, creative, and practical evaluation checklists with as many relevant questions as possible for each user query (one checklist per user query).

You MUST strictly follow the guidelines below:

#### [Guidelines]

- The output should be formatted as a JSON-formatted Python list.
- Each entry in the list should be a Python dictionary containing the following keys: "utterance\_id" and "checklist".
- The "utterance\_id" should indicate the utterance index of the user's utterance in the conversation, with the utterance index always starting at 1.
- Each entry in the "checklist" should be a Python dictionary containing the following keys: "question", "main\_criteria", and "sub\_criteria".
- The "question" should indicate a very detailed and specific question that should be answerable with the positive answer of "Yes", and should be relevant to "main\_criteria" and "sub\_criteria".
- The "main\_criteria" should indicate the primary criteria of the "question", from the given collections.

- The “sub\_criteria” should indicate more fine-grained criteria (represented as a concise noun phrase) belonging to “main\_criteria”.
- In each question of the “checklist,” it is NOT necessary to have exactly one question for each “main\_criteria.” That is, multiple questions can be generated for a single “main\_criteria” as needed.

[Evaluation Checklists]

Only generate the JSON-formatted evaluation checklists without any additional descriptions or explanations.

### Prompt Template used for MULTIVERSE Evaluation: Quality Assessment

You will be provided with an image, a previous dialogue history between the user and the model, a reference answer that gets a score of 10, and an evaluation checklist.

Your task is to evaluate the quality of the model’s answer to the given dialogue history and image based on the provided evaluation checklist, which contains multiple questions. Compare the model’s answer to the reference answer.

### Previous Dialogue History

{dialogue\_history}

### Model’s Answer to evaluate:

{model\_answer}

### Reference Answer (Score 10):

{reference\_answer}

### Checklist (Evaluation Items)

{checklist}

Please use this checklist to guide your evaluation, but do not limit your assessment to it. Compare the model’s answer to the reference answer based on the checklist and the detailed criteria below. Scores should range from 1 to 10, where 1 indicates a very poor response and 10 signifies a perfect response. Here are more detailed criteria for the scores:

[How well does the model perform overall in terms of accuracy, coherence, reasoning, informativeness, and user satisfaction?]

- Score 1-2: The model’s response is generally poor, with major issues in accuracy, coherence, reasoning, and relevance, making it unhelpful.
- Score 3-4: The model’s response is somewhat useful but contains noticeable flaws in accuracy, logical reasoning, or relevance that limit its effectiveness.
- Score 5-6: The model’s response is moderate, demonstrating reasonable accuracy, coherence, and informativeness, though some aspects may be improved.
- Score 7-8: The model’s response is strong, showing high accuracy, logical reasoning, and relevance with only minor weaknesses.
- Score 9-10: The response is excellent, thoroughly accurate, and offers all necessary information to solve the user’s problem.

### Output Format:

Please provide your evaluation results in the following JSON format by filling in the placeholders in

```
:
json
{
  "score": "[1 10]"
}
```

Do not include any additional explanations or descriptions.

Answer:

### Prompt Template used for MULTIVERSE Evaluation: Checklist Completion Accuracy

You will be provided with an image, a previous dialogue history between the user and the model, a reference answer, and an evaluation checklist.

Your task is to evaluate the quality of the model's answer to the given dialogue history and image based on the provided evaluation checklist, which contains multiple questions. Compare the model's answer to the reference answer.

For each question, answer "Yes" or "No."

### Previous Dialogue History  
{dialogue\_history}

### Model's Answer to evaluate:  
{model\_answer}

### Reference Answer (Ground Truth):  
{reference\_answer}

### Checklist (Evaluation Items)  
{checklist}

### Output Format:  
Provide the final answer in the format of "<Q>: <Yes or No>". Do not include any additional explanations or descriptions.

Answer: